Dynamic Key Frame Presentation Techniques for Augmenting Video Browsing

Tony Tse, Gary Marchionini, Wei Ding, Laura Slaughter, Anita Komlodi
Digital Library Research Group
College of Library and Information Services
University of Maryland
College Park, MD 20740, USA
E-mail: {tsetony, march, weid, lauras, komlodi}@oriole.umd.edu

ABSTRACT

Because of unique temporal and spatial properties of video data, different techniques for summarizing videos have been proposed. Key frames extracted directly from video inform users about content without requiring them to view the entire video. As part of ongoing work to develop video browsing interfaces, several interface displays based on key frames were investigated. Variations on dynamic key frame "slide shows" were examined and compared to a static key frame "filmstrip" display. The slide show mechanism displays key frames in rapid succession and is designed to facilitate visual browsing by exploiting human perceptual capabilities. User studies were conducted in a series of three experiments. Key frame display rate, number of simultaneous displays, and user perception were investigated as a function of user performance in object recognition and gist determination tasks. No significant performance degradation was detected at display rates up to 8 key frames per second, but performance degraded significantly at higher rates. Performance on gist determination tasks degraded less severely than performance on object recognition tasks as display rates increased. Furthermore, gist determination performance dropped significantly between three and four simultaneous slide shows in a single display. Users generally preferred key frame filmstrips to dynamic displays, although objective measures of performance were mixed. Implications for visual interface design and further questions for future research are provided.

KEYWORDS: video browsing, representations, dynamic displays, key frames, display rate, divided attention, interface design

INTRODUCTION

Digital video is commonly required by applications such as digital libraries and distance learning. However, the basic characteristics of video—synchronized spatial and temporal coordination of moving images, textual data such as closed-caption, and audio information—raise fundamental issues: What is the basic unit of video? What attributes are best suited to objectively describe its content? What is an acceptable tradeoff level between task completion time and accuracy? While research on such questions is prevalent from a systems perspective, this study focuses on the user perspective.

Users increasingly need to retrieve and manipulate digital video. However, current information retrieval techniques do not provide users with effective mechanisms to review and select video. In addition, there are few interface designs that address how visual information can be obtained rapidly without viewing the whole video. One approach is to design interfaces that facilitate visual browsing and place cost/benefit decisions under user control. In general, browsing is dependent on the availability of appropriate representations and effective user control mechanisms. In this paper, the term "surrogate" refers to representations that people scan and examine to extract meaning and make rapid decisions about further processing. For example, in text-based systems, documents can be represented as titles, bibliographic citations, and abstracts. What are the equivalent surrogates for video? What unique properties of video can be exploited to create surrogates? What are appropriate control mechanisms? Table 1 summarizes various surrogates and mechanisms that bear investigation.

Types of Surrogates

- Key frame (or *poster frame* and *thumbnail*)
- Bibliographic information (title, producer, date)
- Linguistic descriptors
- Linguistic extracts
- Visual extracts (color, luminosity, and subsets or skims)
- Audio extracts (speech, music, and sound effects)

Types of User Control Mechanisms

Static

- Filmstrip (or *storyboard*)
- Labeled thumbnails
- Hierarchical arrangement of surrogates

Dynamic

- Temporally accelerated surrogates (fast forward)
- Multiple parallel surrogates
- Slide show

Table 1. Examples of surrogates and display types for video browsing interfaces.

One of the goals of the Digital Library Research Group is to explore design parameters for user-centered browsing interfaces based on user needs unique to characteristics of a particular data type. Specifically, the group aims to (1) understand how humans process information objects, (2) develop representations of information objects to augment human processing, and (3) explore interactive control mechanisms to enhance human processing.

The exploratory studies presented in this paper focus on *visual* presentation techniques for facilitating or augmenting the user's capability to browse video data (other modalities such as audio and text were not studied). The main question was how effectively users could recognize, retain, and comprehend video using dynamic (i.e., slide show) and static (i.e., filmstrip) surrogates. Specifically, task performance accuracy was measured as a function of viewing rate. A slide show surrogate presents individual key frames consecutively. Two techniques for decreasing overall viewing time were tested: varying the speed of key frame presentation [4] and varying the number of simultaneous slide show displays [12]. In addition, a filmstrip surrogate that displayed four rows of three key frames in temporal order from each video was compared to a dynamic slide show display [8].

Background

Browsing is an effective strategy for information seeking that complements automatic information retrieval techniques, especially for problems in which precise queries cannot be easily formed. A limitation is that it is only practical for a relatively small set of objects [9]. In general, browsing display mechanisms are most useful in information retrieval for scanning ranked search results. A number of video surrogates have been proposed to facilitate visual browsing (Figure 1).

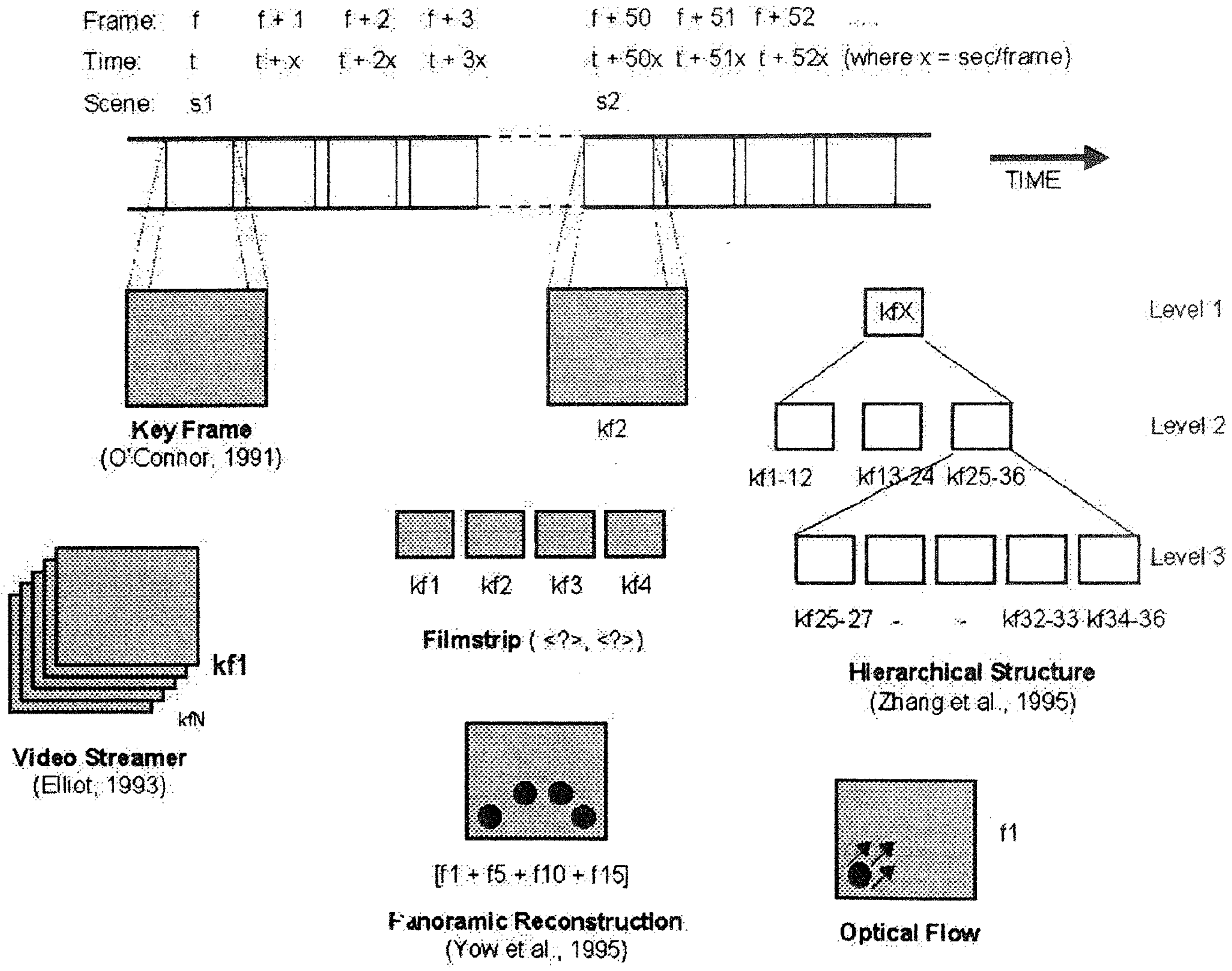


Figure 1. Schematic diagram illustrating different techniques proposed for summarizing video.

O'Connor [10] suggested that key frames, still images representative of scenes extracted from the video itself, could be used to construct video abstracts or "contour maps." Elliot's [6] Video Streamer stacks still images from videos in temporal order, forming three-dimensional "blocks." Viewers can rapidly scan the edges of the blocks to identify both areas of intense motion and types of motion represented. Whereas key frames are surrogates that are extracted directly from the original video, other types of surrogates are synthesized. Yow et al. [17] used "panoramic reconstructions" to summarize high activity scenes from a soccer match. The resulting montage, created by superimposing consecutive images at intervals, clearly illustrates object motion in still images. Similarly, optical flow computation [1] has been used to summarize motion. Object displacement over time is represented by a series of vectors emanating from different objects in the image. For example, Teodosio and Bender [13] used optical flow techniques to derive salient stills, a surrogate that superimposes key images from segments of video into a single frame.

While individual surrogates (e.g., single key frames) can provide some useful information about video, higher-order structures allow for the preservation of temporal as well as compositional information present in the original video. Filmstrip surrogates display multiple key frames in sequence, retaining temporal integrity of the video. Hierarchically ordered key frames [18] are not only temporally ordered, but provide support for changing "resolution." At the top of the hierarchy, a single key frame that best represents the video is shown. Each subsequent level provides greater numbers of key frames displayed as filmstrips, allowing users to navigate progressively to levels where key frames represent individual shots. Thus, the key frames themselves serve as indices, conserving user search time and screen real estate. Yeung et al. [16] used a directed graph display to "cluster" video categories. In the hierarchical scene transition graph model, categories based on overall similarity between shots are shown as nodes and are represented by a single key frame. Edges drawn between the nodes indicate temporal relationships, helping preserve some narrative structure. In the video skim [14], short sequences of consecutive frames are extracted from "important" parts of the video—as measured by scene changes and breaks, audio level, and other cues—and spliced together. Skims are most similar in concept to movie previews, showing only the most salient scenes.

Others, such as Christel, Winkler, and Taylor [3], have conducted user studies on static video surrogates. This study reports user performance and satisfaction results from a series of dynamic display conditions and compares them with static video displays. Dynamic visual surrogates provide certain advantages over static displays. Physically, they require less screen space as each subsequent frame is projected onto the same area. Temporally, they preserve thematic order using the same mechanism as video itself. From the user perspective, dynamic visual surrogates take advantage of the visual systems capability to detect and recognize motion.

Research Questions

Browsing consists of a series of decisions that users make about whether to continue examining a particular information object. That is, relevance in browsing is an affirmation of continued browsing or successful information extraction rather than a final destination. It was hypothesized that dynamic surrogates could facilitate effective browsing of video.

The experiments tested the effects of low-level slide show parameters on user performance in tasks common to browsing. Subjects were assessed on object recognition and gist determination. The following questions were addressed:

- How does key frame display rate affect user performance in object recognition and gist determination?
- How does the number of simultaneous slide show displays affect user performance in object recognition and gist determination?
- How does a second exposure to multiple slide show displays affect user performance?
- How does user performance using static filmstrips compare with performance using dynamic slide show displays?

More generally,

- How can the system provide flexible ways for users to optimize time/benefit tradeoffs (i.e., quick relevance judgements) according to their immediate information needs?

APPROACHES

Static and Dynamic Key Frame Displays

A single key frame surrogate can represent an entire video. For a 3-minute video segment at 30 frames per second, this represents an information *compaction (IC) ratio* of 5400:1. If the single key frame requires one second to browse, it would have a *viewing compaction (VC)* ratio of 180:1. That is, by using a single key frame to represent the 3-minute segment, a viewer would save 179 seconds or 179/180 of the time required to view the video in real time.

IC ratio = video length (s) * video rate (fps) / number of key frames (f)

VC ratio = video length (s) / browse time (s)

To provide more information to the user, one key frame may be selected from each shot. For example, if the 3-minute segment consisted of 30 shots, 30 key frames would represent that segment (IC ratio of 180:1). Assuming that a viewer requires 1 second per key frame to absorb the visual information contained in the images, she or he would require 30 seconds to view all of them (VC ratio of 6:1). From a user-centered perspective, VC ratio determines the efficacy of video browsing.

Screen space is an issue with typical static key frame displays. For example, a filmstrip display is typically formatted so that key frames are displayed from left to right and top to bottom. Thus, in addition to the time required for scanning each key frame in a row, at the end of each row the eyes must traverse the screen to the beginning of the next row. Furthermore, if all of the key frames cannot fit on a single screen, additional time is required for clicking or scrolling. After each of these actions is taken into account, the effective VC ratio decreases and the compaction advantage diminishes. Additionally, the surrogate browsing mechanism will in practice be embedded in a larger process (e.g., examining a series of video clips retrieved as a result of a query). The screen real estate needed for surrogate display may occlude the larger task elements and also slow overall task performance.

Both the time required for eye motion and user interface activities in static frame displays may be eliminated through a dynamic slide show mechanism. In this arrangement, only a single key frame is visible on the screen at any time and the size of the surrogate can be adjusted for optimal viewing. Thus, the viewer's eyes can dwell in the same region of the screen and there is no need for scrolling or other user action.

One interesting consequence of dynamic key frame displays is that increasing the rate at which key frames are shown can increase the VC ratio without modifying the number of key frames. For example, with 30 key frames extracted from a 3-minute video segment, the VC ratio jumps from 6:1 to 12:1 as the rate is doubled from 1 key frame per second (kfps) to 2 kfps. If people can effectively browse key frames in less than one second, the VC ratio can be greatly improved. Under certain conditions, visual processing for understanding images can be performed in 100 ms or less [11]. Thus, for specific video browsing tasks, a slide show presentation of 30 key frames shown at 10 kfps (a VC ratio of 60:1) may be achieved.

Another approach to increasing VC ratios is displaying several slide show displays simultaneously. Although it may be possible for a viewer to attend to several slide shows of key frames at the same time, a decrease in user performance, for example, based on Wickens' [15] multiple resource theory, would not be unexpected. Two simultaneous displays on the screen would increase the VC ratio by a factor of two. That is, if a single display provides a VC ratio of 6:1, two simultaneous displays would increase it to 12:1.

Metrics

In all three of the experiments, subjects were requested to participate in two general visual browsing activities: object recognition and gist determination. One of the experiments additionally used an action recognition metric. The tasks were selected to represent different user needs and browsing strategies.

Object Recognition

The object recognition (OR) task was used to simulate situations in which users browse unstructured environments with fuzzy ideas of what they need. For example, a biology teacher might browse video surrogates, looking for scenes that illustrate the effects of pollution on the environment. Object classes might include industrial waste containers, smoke plumes, and affected wildlife. Because the teacher may wish to review all the key frames in the video surrogates before making a decision as to whether or not a particular scene should be reviewed in more detail, recognized objects must be retained in memory for a short period of time. This task is differentiated from object identification, in which specific instances of an object (e.g., the Chernobyl nuclear power plant) are sought rather than a general category (e.g., industrial complexes). The environment is unstructured in that key frames represent entire scenes and so viewers cannot easily anticipate where objects will appear in any given frame.

The OR task was operationalized through cued recall. Alphabetical lists of 20 objects, half of which actually appear in the key frames (targets) and half of which do not appear at all, but are contextually consistent with the targets and overall theme of the video (distractors), were created for each condition. Subjects were shown the list for a short period prior to viewing a slide show display. After viewing a display, subjects were asked to indicate all objects on the list that appeared in at least one key frame. There was no time limit. In assessing task performance, one point was given for each target object selected and each distractor object *not* selected. Total points accumulated for each list were used as a measure of object recognition performance. Thus, scores could range from 20 (perfect performance) to 0 (worst performance). Cued recall using targets and distractors is common in visual processing and attention research. However, most of these tests have used both controlled environments and objects drawn to scale with distinct characteristics (e.g., computer-generated geometric shapes). Because this study used still images extracted from actual videos, determining target objects that were "similar" in visibility [e.g., size, luminosity, relative depth (foreground vs. background)] and creating "reasonable" distractors was difficult. However, several iterations of list refinement and pilot testing seem to support the face validity of this metric.

Gist Determination

The gist determination (GD) task was designed to determine how much thematic information subjects could obtain from browsing video key frames. Whereas the object recognition task focused on whether particular object types appear in a particular set of key frames, this task looks at how well users can determine the overall meaning of a video from viewing only the key frames.

Two different methods were used to measure gist determination performance: free form sentence creation and multiple choice statement selection. In the first method, subjects were requested to write a brief summary describing what they thought the theme of the actual video was after viewing the video surrogates. This technique allows for greater expressiveness in the responses, resulting in greater variability. The second method, providing subjects with four statements on the theme of the video and asking them to select the "best" one (i.e., three distractors and one target), results in less variability but limits the amount of detail that can be recorded. There were no time limits. GD performance was estimated with both methods. In sentence creation, content analysis was conducted. In the rate experiment, subject responses were placed into three categories based on depth of comprehension and degree of involvement of external knowledge: correct literal objects or events (one point), correct general thematic information or "common sense" judgments (two points), and accurate thematic information (three points). In the simultaneous display experiment, they were categorized into four topics: people, objects, actions/concepts, and places. In statement selection, identification of the target statement was given a single point, while selection of a distractor sentence resulted in zero points.

GD could be triangulated between the two methods—a correct, highly conceptual original sentence with selection of the target statement by a single subject is more likely to indicate higher performance than an incorrect statement of fact and selection of a distractor statement. However, inherent problems in this metric include too few choices in the statement selection (a one-in-four or 25 percent chance of selecting the correct statement at random) and difficulty in assessing true understanding of meaning from a single writing sample. For example, the writing task introduces a bias against subjects who do not express themselves well in written form. Furthermore, background knowledge or previous experience of the subject matter shown in the slide show displays was not controlled.

Action Recognition

In one of the experiments, the action recognition (AR) metric was used when evaluating sentence creation as an "intermediate" measure between the OR and GD tasks. Whereas OR attempts to measure types of objects recognized, the AR task was designed to test subjects' ability to identify and relate multiple objects to general actions without needing to fully understand the context. Thus, the cognitive task measured by AR requires more cognitive effort than OR but less than GD.

The AR task was operationalized by free-form sentence creation. Subjects were asked to write sentences describing what they saw. A point was given for each correct description of an action and a point was deducted for each "incorrect" description.

User Perception

Questionnaires were used to obtain perception data. In the variable rate experiment, user perception of the rates tested was measured using a seven point Likert scale from slow (1) to fast (7), with a score of 4 indicating "neither." Subjects were asked about speed perception for both the OR and GD tasks. It was expected that GD would be rated closer to the middle than OR, as determining overall meaning was hypothesized to require less cognitive effort than remembering whether specific object types were encountered. For the variable simultaneous display experiment, the same Likert scale was used to rate display speed for both OR and GD. In addition, subjects were asked about a number of simultaneous screens on a seven-point Likert scale, from imperceptible (1) to perceivable (7). For the experiment comparing performance between dynamic and static slide show displays, subjects were asked to write evaluative sentences halfway through the session and at the end of the session.

User perception is an especially important metric since it is the only way to get at subjective measures. Although the measures are mainly qualitative, user comments can provide designers with invaluable feedback. Different conditions can be compared to each other. Not only can triangulation of perception scores with quantitative data validate trends, significant differences between perception and performance data can indicate that user preference does not correlate with performance. Such seemingly incongruous results can provide a useful insight into human factors and provide more information for developing more accurate cognitive models.

METHODS

Overall video browsing effectiveness is expected to vary with factors such as user characteristics, tasks to be completed, and tools available. For example, a user with experience in analytical text-based searching who needs to find a particular video object with a known title and director is likely to benefit more from a traditional keyword-based user interface than a dynamic key frame display. As noted previously, browsing is an information seeking strategy that helps users who do not have a clear goal make decisions by rapidly rejecting records not of interest. This interactive filtering process allows users to narrow a result set of records and find the items that satisfy their needs [9]. The different dynamic presentations described in this study are designed to support precisely this aspect of browsing.

Because of the potentially large number of variables that could affect browsing effectiveness, specific factors were tested, while many others had to be ignored or excluded in the current set of experiments. With respect to user characteristics, although demographic data were collected and analyzed, other relevant variables such as previous experience with video editing, user fatigue, and spatial visual ability were not controlled. Tasks studied across all three experiments were limited to object recognition and gist determination. Finally, system variables such as screen size, viewing distance, viewing resolution, and network versus local access to various test files were not controlled.

Similar procedures were used for each of the three experiments: key frame rate, number of simultaneous displays, and comparison of dynamic and static displays (Table 2). All subjects were briefed on the prototype interface and given a demonstration. Following a practice trial, a series of experimental trials were conducted. Subjects were then asked to complete a variety of questionnaires, including those for demographic information, and debriefed.

Key Frame Rate Experiment

One practice session and five experimental sessions (varying frame rates) were administered to each subject. Subjects were tested individually in the office of one of the authors. Subjects were shown a list of 20 objects prior to each session. After viewing each slide show (Figure 2, next page), they were asked to complete three tasks in writing: object recognition, gist determination, and user perception. After the sixth and final session, subjects were debriefed. Both the video segments and display rates were randomized for each subject to minimize display rate and order biases. Subjects were only allowed to view each slide show presentation once. (See [4] for details.)

Simultaneous Displays Experiment

One practice session and two experimental sessions were administered to each subject. Subjects were tested in groups at the University of Maryland's Academic Information Technology Services (AITs) Teaching Theater. During the practice session, a single slide show was shown. The experimental sessions immediately followed (see Figure 3, next page). Subjects were requested to complete three tasks on-line after each test session: object recognition, gist determination, and user perception. Each subject had been randomly assigned to view one of the four experimental display conditions (one display, two, three, or four simultaneous displays). Subjects were only allowed to view the display once. (See [12] for details.)

	Key Frame Rate	Simultaneous Displays	Comparison: Dynamic & Static
Subjects (U of Maryland students)	20 graduate and undergraduate volunteers	28 introductory psychology undergraduates – for credit	30 psychology major undergraduates – for credit
Source of key frames	Discovery Channel documentaries	Discovery Channel documentaries	Discovery Channel documentaries
Extraction program	Merit[1]	Merit[1]	Merit[1]
Video segments	6	5	7
Screen resolution (pixels / colors)	640 x 480 / 256	800 x 600 / 256	1024 x 768 / 256
Image format	GIF (352 x 240)	GIF (180 x 120)	GIF (180 x 120)
Platform	Apple/Power Mac	Windows/IBM-PC	Windows/IBM-PC
Interface	Netscape Navigator	Netscape Navigator	Netscape Navigator
Display rate(s) (key frames / sec)	1, 4, 8, 12, 16	1	4
Simultaneous dynamic displays	1	1 – 4	1
Static displays	0	0	2
Experimental measurement	Repeated	Completely randomized design	Repeated
Number of treatments	5	4	2
Tasks[2]	GD, OR, UP	GD, OR, UP	GD, AR, OR, UP

[1] Developed at the University of Maryland Center for Automation Research (CfAR)

[2] GD = gist determination; AR = action recognition; OR = object recognition; UP = user perception

Table 2. Summary information comparing methodology among three experiments.

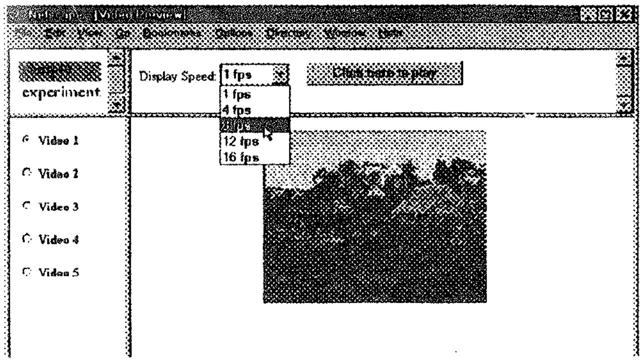
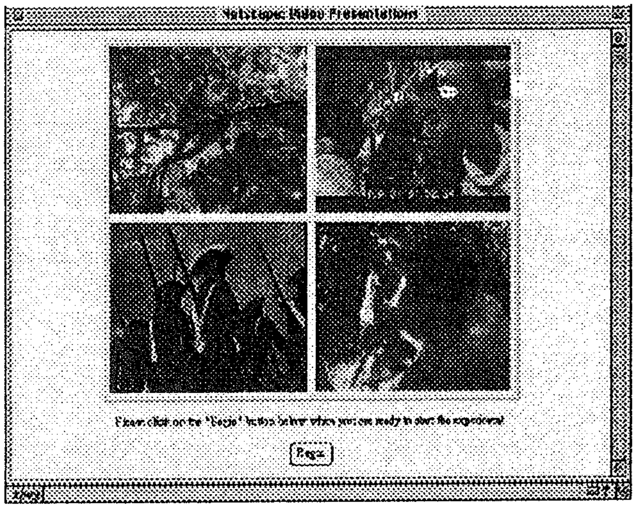


Figure 2. Key frame rate experimental interface.
Pictures shown are Copyright © Discovery Channel, Inc.
All rights reserved.

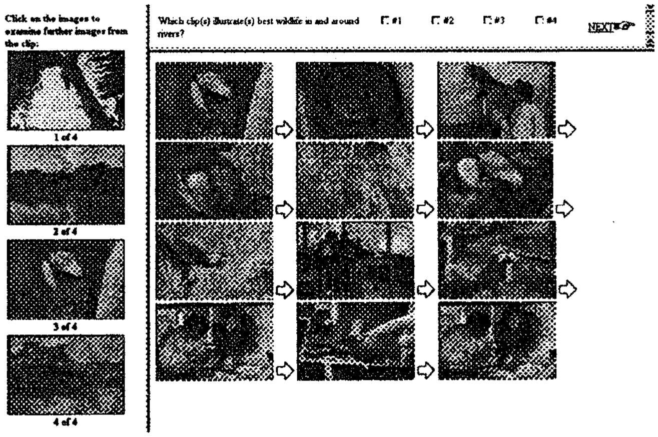


Please click on the "Begin" button below when you are ready to start the experiment.

Figure 3. Simultaneous displays experimental interface (4 displays shown).
Pictures shown are Copyright © Discovery Channel, Inc.
All rights reserved.

Dynamic Versus Static Displays Experiment

Four experimental sessions were conducted for each subject. In the first three sessions, subjects were shown three video objects arranged either as static 12-key frame filmstrip displays, static 4-key frame filmstrip displays, or dynamic slideshow displays (subjects were randomly assigned to one of the three treatment groups). There was no time limit on viewing time for the static displays and subjects were allowed to view the slide show display as many times as desired. Subjects were asked to complete three tasks on-line immediately following each session: identifying objects from a list, describing what they saw in the clips in sentences, and selecting one-sentence descriptions about the clips (multiple choice). User perception information was collected after the three sessions. In the last session, subjects were shown four video surrogates representing different video objects ("compressed format") in a static display (see Figure 4). For more detailed browsing, clicking on any of the four single key frames caused 12 additional key frames in filmstrip format from the same video object to be displayed in an "expanded format." There was no time limit on browsing key frames in either the compressed or expanded format. In contrast to the other experiments in which task performance was assessed after the video display was no longer visible, in this session subjects were shown questions on the screen simultaneously with the video display. Thus, rather than relying on short-term memory as the other metrics do, this protocol was an attempt to simulate a realistic information seeking environment in which specific questions are known by users *during* browsing. User satisfaction was tested after this session. (See [8] for details.)



Click on the images to examine further images from the clip:

Which clip(s) illustrate(s) best wildlife in and around rivers? □ #1 □ #2 □ #3 □ #4

Figure 4. Dynamic vs. static experimental interface (static display shown).
Pictures shown are Copyright © Discovery Channel, Inc.
All rights reserved.

RESULTS

The results from the three experiments together contribute to an overall understanding of how dynamic slide show displays may be useful for specific types of video browsing tasks. Both quantitative and qualitative data were obtained.

Key Frame Rate Experiment

Overall, performance in object recognition (OR) decreased with increased key frame rates (Figure 5, next page). Statistically significant differences between performance and key frame display rate were found for three homogeneous subsets: 12 and 16 kfps, 8 and 4 kfps, and 1 kfps (one-way ANOVA, F(4, 65) = 12.35, $p<0.00$). Statistical significance at the 0.05 level was found in performance, pairwise, between all five rates (1, 4, 8, 12, 16 kfps) except for two pairs: (1) 4 and 8 kfps and (2) 12 and 16 kfps.

Analysis of gist determination (GD) performance as measured by sentence selection at different display rates indicated no statistically significant differences (one-way ANOVA, F(4, 65) = 0.981, $p<0.4245$). Similarly, no statistically significant differences were found from sentence analysis (one-way ANOVA, F(4, 95) = 2.314, $p<0.0630$).

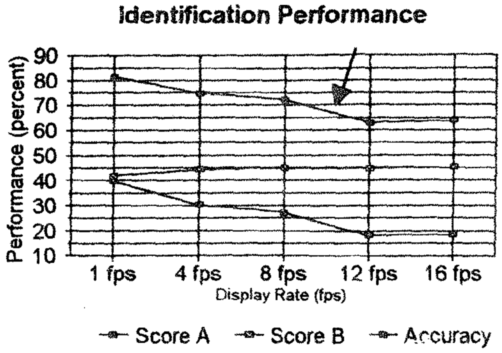


Figure 5. Performance in the object recognition task as a function of display rate (key frames per second).
[Note: Score A = % targets identified correctly; Score B = % distractors identified correctly; Accuracy = % Sum (A + B)]

User perception for OR and GD tasks was compared at each display rate and found to be statistically significant (t-test, p = 0.00) for all rates. In other words, users consistently felt that GD was easier than OR for a given display rate.

Simultaneous Displays Experiment

Overall, object recognition (OR) decreased with number of simultaneous displays viewed. Statistically significant differences in OR performance for one to four slide show displays at 1 kfps were detected using the Kruskal-Wallis non-parametric test for the practice session (H(3) = 15.96, p<0.001) and for the experimental session (H(3) = 12.74, p<0.005).

No significant differences were found for gist determination (GD) performance as a function of number of simultaneous screens. Performance in sentence selection did not vary across experimental conditions: Kruskal-Wallis showed H(3) = 3.88, p<0.25 for the practice session and H(3) = 0.591, p<0.90 for the experimental session. However, qualitative sentence analysis on the results of the sentence creation task revealed a negative relationship between gist understanding and the number of simultaneous screens viewed (Table 3).

Condition	% Correct	# Times Main Idea Unclear
1 Slide Show	87.5	0
2 Simultaneous Slide Shows	73	2
3 Simultaneous Slide Shows	44	4
4 Simultaneous Slide Shows	28	5

Table 3. Analysis of gist determination (n = 7 for each condition).
[Note: % correct = correct concepts identified by subjects as a percentage of all concepts; # times main idea unclear = number of times subjects indicated that they were not sure of the gist and could not give a clearly articulated response]

User perception data showed that subjects perceived simultaneous multiple slide show displays as being "too fast" relative to the single slide show situation. In addition, user perception during the OR and GD tasks indicated that number of simultaneous displays were proportional to amount of "imperceptibility."

Dynamic Versus Static Displays Experiment

Object recognition (OR) performance among the two 12-key frame conditions (i.e., static 12-key frame filmstrip display and dynamic slide show display at 4 kfps) were found to be statistically significantly different (one-way ANOVA, F(1, 18) = 20.743, p<0.00). Performance was better in the static 12-frame condition than with the dynamic key-frame slide show (Figure 6).

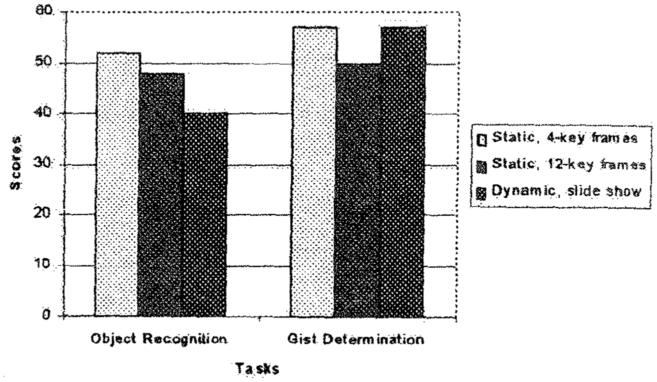


Figure 6. Object recognition and gist determination task scores for 3 key frame display types.

Gist determination (GD) measured by sentence selection showed no statistically significant differences. GD performance based on sentence creation showed no statistically significant differences (p = 0.895). Qualitative content analysis showed that, "in general, subjects tended to identify actions for the beginning of the clips while the second parts were usually described by listing objects. (p. 15)" [8].

Although there were no statistically significant results (p = 0.939), subjects using the 12-key frame static display did best in action recognition (AR) component of the sentence creation task. Subjects using the 4-key frame static display did better than those using the dynamic display.

Even though the GD tasks showed no statistically significant differences, subjects in the dynamic display treatment group scored the highest in the sentence writing tasks and scored the same as those in the 4-key frame static display group in the multiple sentence assessment. The 12-key frame static display group had the lowest scores in both tasks.

Quantitative analysis of user perception indicated no statistically significant differences in user satisfaction but did show a significant difference in user satisfaction ratings between single and multiple static video displays ($p<0.00$). Subjects preferred multiple displays because they could view additional surrogates from four clips on a single screen and directly control which clip to view in detail.

CONCLUSIONS AND RECOMMENDATIONS

The experiments described in this paper focus on (1) usability as a function of two different parameters, key frame rate and number of simultaneous displays, and (2) user perception, for specific tasks (i.e., object recognition, action recognition, and gist determination) using different key frame slide show display mechanisms. Although the ultimate goal of this research is to understand how visual displays can support and augment user needs for browsing video, the current research is exploratory and limited in scope. However, the results provide interesting insight into techniques for augmenting user-centered video browsing.

The results showed that, in general, users could recognize objects without significant performance degradation (-10%) at relatively high key frame rates (up to 8 kfps). This is consistent with previous results, [7] and [11], suggesting that the time required for visual object recognition is about 100 ms or the length of a single saccade (8 kfps is equivalent to 125 ms/key frame). Furthermore, subjects could reasonably recognize objects from up to three simultaneous slide show displays, suggesting that visual information processing is limited by some psychophysiological bottleneck (e.g., multiple resources theory [15]). The tests were conducted on subjects who represent the occasional user. An interesting question is whether prolonged exposure to high levels of simultaneous, rapid visual input could increase an individual user's capacity to process visual information. Although some demographic information such as amount of time spent watching television was collected in an attempt to understand how longer-term experience might affect visual throughput, no significant differences were found in these limited data. Another factor might be the effect of learning specific to these interfaces. Finally, users' native capabilities, such as visual or spatial abilities, may influence the limitations found in these experiments. It is thus likely that no single visual-processing "threshold" exists for users, but rather a range of effective levels based on many different factors.

The results reported here suggest that gist determination is influenced by visual information throughput. The greatest performance difference in this task was found to occur between 8 and 12 kfps and between two and three simultaneous slide show displays at 1 kfps (or virtually 2 and 3 kfps, respectively). The cognitive process of linking disparate information (objects) into coherent frameworks, or schemas [5] may partially explain these results. It is hypothesized that while browsing key frames, users create schemas that are consistent with the objects recognized both within a single key frame and among key frames through time to derive meaning. It is this construct that is used to complete the gist determination task. However, if objects have been misidentified or missed completely due to information overload, then incorrect (e.g., misinterpretations) or incomplete (e.g., object-level sentence construction or user uncertainty in the gist determination task) schemas are formed. This may partially explain the inverse relationship between information density (increasing key frame rates and/or number of simultaneous displays) and accurate gist determination.

User perception is very important. An efficient interface that causes users cognitive discomfort or rapid mental fatigue is problematic. The results of these experiments suggest that even though users were capable of performing within "acceptable" ranges at high viewing compaction (VC) ratios, their self-reported subjective states indicated that they felt "overwhelmed" by the high level of data throughput. Part of this may be due to a "novelty effect"; although rapidly changing visual stimuli have become commonplace in popular culture (e.g., television, the Web, and advertising in general), the goal of these dynamic images isn't to inform per se, but rather to entertain or influence. Perhaps users find these stimuli easier to process, not because of any significant differences in visual information density, but rather the complex experimental tasks that they are required to complete in the experimental condition. Another explanation is that users' discordant responses to satisfaction questions may result from their inability to control the interface. In an actual interface prototype, of course, users would be given controls to determine the display rate or number of displays to view at the same time. But in order to determine how performance is affected, the parameters were set by the investigators (only one of the experimental interfaces allowed for minimal user control). Future experiments are needed to study how interactive control mechanisms affect user satisfaction.

There are still many areas to be addressed. From the user side, performance while varying rate and number of displays as a function of age, gender, educational background, and other demographic factors needs to be studied. Furthermore, cognitive abilities such as spatial-visual abilities, background knowledge of the subject matter, cultural experience, and other individual factors need to be controlled. In addition, physiological factors such as visual fatigue need to be examined.

From a task perspective, new and improved methods for assessing performance are required. The metrics reported in these experiments are but a first cut in trying to achieve face validity in measuring object recognition and gist determination. As implemented, subjects might be selecting words through association rather than from their visual cues from working memory in the object recognition task. One way to test for this, for example, is to employ eye-tracking apparatus to determine if subjects had actually foveated on a correctly identified target object. If so, then there is stronger evidence that the subject did, in fact, recognize the object based on visual input rather than using intuition or deductive logic. Another way to test target object validity would be to create highly controlled scenes through computer graphics (e.g., [2]). However, such conditions would not be as realistic as using key frames from actual video objects. Also, the object recognition and gist determination tasks tested in these experiments represent narrow behaviors in the spectrum of user video browsing needs. Between these extremes is a variety of other information seeking behavior that plays at least as great a role. One of the experiments attempted to create situations that more closely reflect users' needs. More research on actual user needs would inform interface developers which features are most likely to help augment user video browsing.

Finally, from a system perspective, various presentation conditions (e.g., screen resolution, visual angle, and use of colors) need to be studied further. For example, colors and motion can be detected more readily in peripheral vision. A more fundamental question alluded to earlier, whether key frames actually are optimal as video surrogates, needs to be studied. A number of other techniques, both static and dynamic, have been proposed. These should be researched in usability testing laboratories to determine performance range and compared with each other. Only in this way can a more robust theory and practice of creating video surrogates be developed and implemented.

In summary, it is clear that different presentation display techniques are effective under different conditions. Factors to be considered include user characteristics, subject domain, and task. Under certain conditions, dynamic key frame displays seem to augment user browsing effectiveness while static key frame presentations appear to be more effective in other situations. Future studies will need to concentrate on defining how different factors affect the selection of interface style and creating designs that enable users to adapt the video browsing interface to their own abilities, tasks, and preferences.

ACKNOWLEDGEMENT

This work was partially supported by the U.S. Department of Education Technology Challenge Grant #R303A50051.

REFERENCES

1. Beauchemin, S.S. and J.L. Barron. (1995). The computation of optical flow. *ACM Computing Surveys*, **27**(3): 433-466.
2. Boyce, S.J., A. Pollatsek, and K. Rayner. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, **15**(3): 556-566.
3. Christel, M.G., D.B. Winkler, and C.R. Taylor. (1997). Improving access to a digital video library. In *Human-Computer Interaction: INTERACT97*, Sydney Australia.
4. Ding, W., G. Marchionini, and T. Tse. (1997). Previewing video data: Browsing key frames at high rates using a video slide show interface. *Proceedings of the International Symposium on Research, Development, and Practice in Digital Libraries (ISDL '97)*, Tsukuba, Japan.
5. Ellis, H.C. and R.R. Hunt. (1989). *Fundamentals of human memory and cognition.* Wm. C. Brown Publishers: Dubuque, IA.
6. Elliot, E. (1993). Watch, grab, arrange, see: Thinking with motion images via streams and collages. MSVS Thesis Document. Cambridge, MA: MIT Media Lab.
7. Healey, C.G., K.S. Booth, and J.T. Enns. (1996). High-speed visual estimation using pre-attentive processing. *ACM Transactions on Computer-Human Interaction*, **3**(2), 107-135.
8. Komlodi, A. (1997). Visual surrogates for motion picture documents: Presentation techniques for key frames. CLIS-TR-97-15, College Park, MD: Digital Library Research Group (*http://www.glue.umd.edu/~dlrg/*).
9. Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge Series on Human Computer Interaction. Cambridge University Press: New York.
10. O'Connor, B.C. (1991). Selecting key frames of moving image documents: A digital environment for analysis and navigation. *Microcomputers for Information Management*, **8**(2), 119-133.
11. Potter, M.C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory*, **2**(5), 509-522.
12. Slaughter, L., B. Shneiderman, and G. Marchionini. (1997). Comprehension and object recognition capabilities for presentations of simultaneous video key frame surrogates. In Peters C. and C. Thanos (Eds.), *Research and Advanced Technology for Digital Libraries: Proceedings of the First European Conference* (pp. 41-54). ECDL'97, Pisa, Italy.
13. Teodosio, L. and W. Bender. (1993). Salient stills from video. *Proceedings of ACM Mulitmedia '93*, Anaheim, CA: 39-46.
14. Watclar, H.D., T. Kanade, M.A. Smith, and S.M. Stevens. (1996). Intelligent access to digital video: Informedia project. *Computer*, **29**(5), 46-52.
15. Wickens, C.D. (1992). *Engineering psychology and human performance*. Second Edition. New York: HarperCollins.
16. Yeung, M.M., B.L. Yeo, W. Wolf, and B. Liu. (1995). Video browsing using clustering and scene transition on compressed sequences. *Proceedings of Multimedia Computing and Networking*, San Jose.
17. Yow, D, B.L. Yeo, M.M. Yeung, and B. Liu. (1995). Analysis and presentation of soccer highlights from digital video. *Proceedings of the Second Asian Conference on Computer Vision* (ACCV '95).
18. Zhang, H.J., C.Y. Low, and S.W. Smoliar. (1995). Video parsing and browsing using compressed data. *Multimedia Tools and Applications*, **1**, 89-111.